Standards Architecture for the Pathogen Genomic Surveillance Enterprise:

Workshop Learnings and Outcomes

Marc Salit & Arend Sidow, Viridae, Inc. in partnership with The Rockefeller Foundation
Consultation with CSWG & SPHERES on Standards Architecture for Genomic Surveillance

- Pathogen Genomic Surveillance working definition:
  - integrative analysis of viral whole-genome sequencing from individual human-derived samples

- Vision of Standards Architecture
- RF Workshop Takeaways
Why a Standards Architecture?

• Reliable surveillance needs systematic ‘enterprise’
  • robust, integrated, interoperable, verifiable, trustworthy, and reliable

• Enterprise is composed of multiple distinct operations
  • conducted by multiple organizational entities

• “Standards Architecture” is the full suite of standards to enable systematic operation
What are the roles of standards in enterprise performance?

Provide a way to evaluate performance
- “how well can I trust these data?”
- what’s the confidence interval?

Provide the basis of interoperability
- “I can share these data...” and someone else can use them, or compare to them
Invited community to consider **Standards Architecture**

- Shared and reviewed ‘strawperson’ in 4 open, public working group meetings
- Heard substantive feedback, recommendations, discussion of key matters in each group
Diverse Working Group Engagement

- Public/Private/Academic/Non-profit
- Public Health labs, research institutions, clinical labs, technology developers, professional societies, funders
- Laboratorians, informaticians, tool makers, scientists, epidemiologists, standards developers
Pathogen Genomic Surveillance Process Model

**Sampling**

- **Process Component**: Real-Time Epidemiology
- **Flow**: Sampling strategy
- **Example Metadata**: Selection criteria, Other samples in same study, Use case
- **Notes**: The choice of which samples to sequence is driven primarily by the present tense questions

**Individual Sample Process Steps**

- **Components**:
  - Sample Selection
  - Sample Collection and PCR Testing
  - Nucleic Acid Sample
  - Sample Conversion
  - Raw Sequence Data
  - Bioinformatic Analyses
  - Sequence Assembly and Variants
- **Repositories**:
  - Data integration and access
  - Database / Repository
- **Fulfill Distinct Functions**:
  - Support analytical approaches that are still evolving

**Sequence Bioinformatics**

- **Components**:
  - Bioinformatic Analyses
  - Sequence Assembly and Variants
- **Repositories**:
  - Data integration and access
  - Database / Repository

**Integrative Analytics and Archiving**

- **Components**:
  - Bioinformatic Analyses
  - Sequence Assembly and Variants
- **Repositories**:
  - Data integration and access
  - Database / Repository

**Public Health "Use Cases"**

- **Outbreak questions - present tense local**
  - Which strain is responsible for the outbreak?
  - Is there a signal to declare a new VOC?
  - How do we keep it from spreading?
- **Diagnostics and vaccination evolution - present perfect national and global**
  - Have new variants affected PCR-based tests?
  - Do protocols need to evolve?
  - What is the rate of vaccine escape by variants?
- **Science questions - past tense international**
  - How, why, where did the virus spread?
  - What were the origins of the virus?
  - Where did new variants appear?
- **Diagnostics and vaccination evolution - present perfect national and global**
  - Have new variants affected PCR-based tests?
  - Do protocols need to evolve?
  - What is the rate of vaccine escape by variants?
- **Science questions - past tense international**
  - How, why, where did the virus spread?
  - What were the origins of the virus?
  - Where did new variants appear?
  - How fast was their spread?
  - And many questions around function, biology?
Overarching Takeaway I:

Implications of Diverse Use Cases

• **Diversity of Use Cases** drives diversity of Sampling Strategy, Sharing, Analysis
  - global questions of strain prevalence, spread of variants
  - local questions of outbreak identification
  - novel variant / strain detection
  - relevance/willingness to share data

• Takeaways
  - **Metadata Annotation** – one size will not fit all
  - **Standards Development** should be cognizant of critical use cases
Overarching Takeaway II:

Implications of *Diverse Jurisdictions*

- Local, regional, national, international legal and cultural contexts create different requirements and constraints for
  - data sharing (ownership issues)
  - metadata annotation (privacy issues)

- Takeaway
  - Standards Development needs to accommodate jurisdictional contexts
Overarching Takeaway III:

**Diversity of Sample Handling and Measurement**

- Wide variety of methods and technologies
  - processing samples, sequencing, and sequence bioinformatics
- Takeaways
  - need technology-agnostic standards
  - need repository/library of validated tools, protocols, pipelines
  - need authoritatively characterized Reference Materials and Data
    - for optimization and validation of methods and analysis
    - for benchmarking and proficiency testing
Overarching Takeaway IV:

*Metadata* annotation, aggregation, integration is hard

- Metadata annotation is costly
  - not always budgeted
  - burden carried by "upstream" labs, not those analyzing

- Make and deploy better tools
  - tooling to annotate is immature
  - metadata sharing from lab-to-lab or process-to-process is hard
    - e.g., interoperability of *Lab Information Systems* and *Electronic Health Records* is uncommon
Overarching Takeaway V:

*Pathogen Genomic Surveillance* is an evolving discipline

- Pathogen Genomic Surveillance ecosystem is evolving
  - sample handling, sequencing, sequence bioinformatics more mature than epidemiological analyses

- Standards, shared resources, and the *Standards Architecture*
  - must be designed to evolve
  - must have continued engineering to be maintained

- *Can’t hide behind this and eschew rigor*
Potential Standards

- Set of widely available, authoritatively characterized reference samples
- Reference data for sequencing and sequence bioinformatics benchmarking
- Open, public, secure benchmarking platform
- Widely adopted metadata standards
- Benchmarking challenge for integrative analysis